Advanced Factor Analysis Modeling of Spatial Data

P.K. Hopke, P. Paatero, and

J. Hoppenstock

Data

What are the $PM_{2.5}$ mass concentration data as viewed by factor analysts?

It is just a data table (matrix).

DAYVAR	s010030010	s010270001	s010331002	s010491003	s010530002	s010550010	s010690002
990101	-99	-99	-99	-99	-99	-99	-99
990102	-99	-99	-99	-99	-99	-99	-99
990103	-99	-99	-99	-99	-99	-99	-99
990104	-99	-99	-99	-99	-99	-99	-99
990105	-99	-99	-99	-99	-99	-99	-99
990106	-99	-99	-99	-99	-99	-99	-99
990107	-99	-99	-99	-99	-99	-99	-99
990108	-99	-99	-99	-99	-99	-99	-99
990109	-99	-99	-99	**11**	-99	-99	8.7
990110	-99	-99	-99	-99	-99	-99	-99
990111	-99	-99	-99	-99	-99	-99	-99
990112	-99	8.8	-99	-99	-99	-99	14.9
990113	-99	-99	-99	-99	-99	-99	-99
990114	-99	-99	-99	-99	-99	-99	-99
990115	-99	14.9	-99	7.6	-99	-99	-99

Thus, we can look at spatial data as a problem to find the underlying structure that governs the observed values.

- This assumes that the influence of the causal factors extend over the area in which measurements are being made.
- Problem is the same as in weather forecasting where the aim is to predict air mass movement and is known as Empirical Orthogonal Function Analysis

WHY?

- Such analyses can provide a better understanding of the data
- Allow us to predict for specific times and locations for which we do not have data
- Permit network design that provides same information with less data (fewer measurements.)

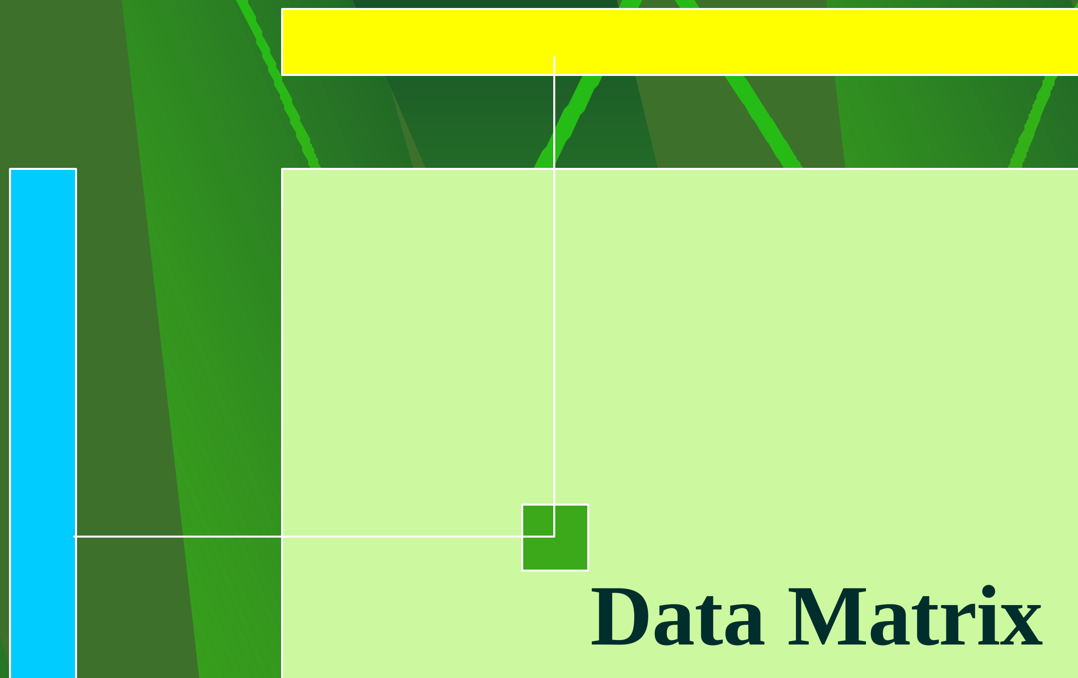
HOW?

- Decompose the matrix into an outer product of vectors
- Huh?

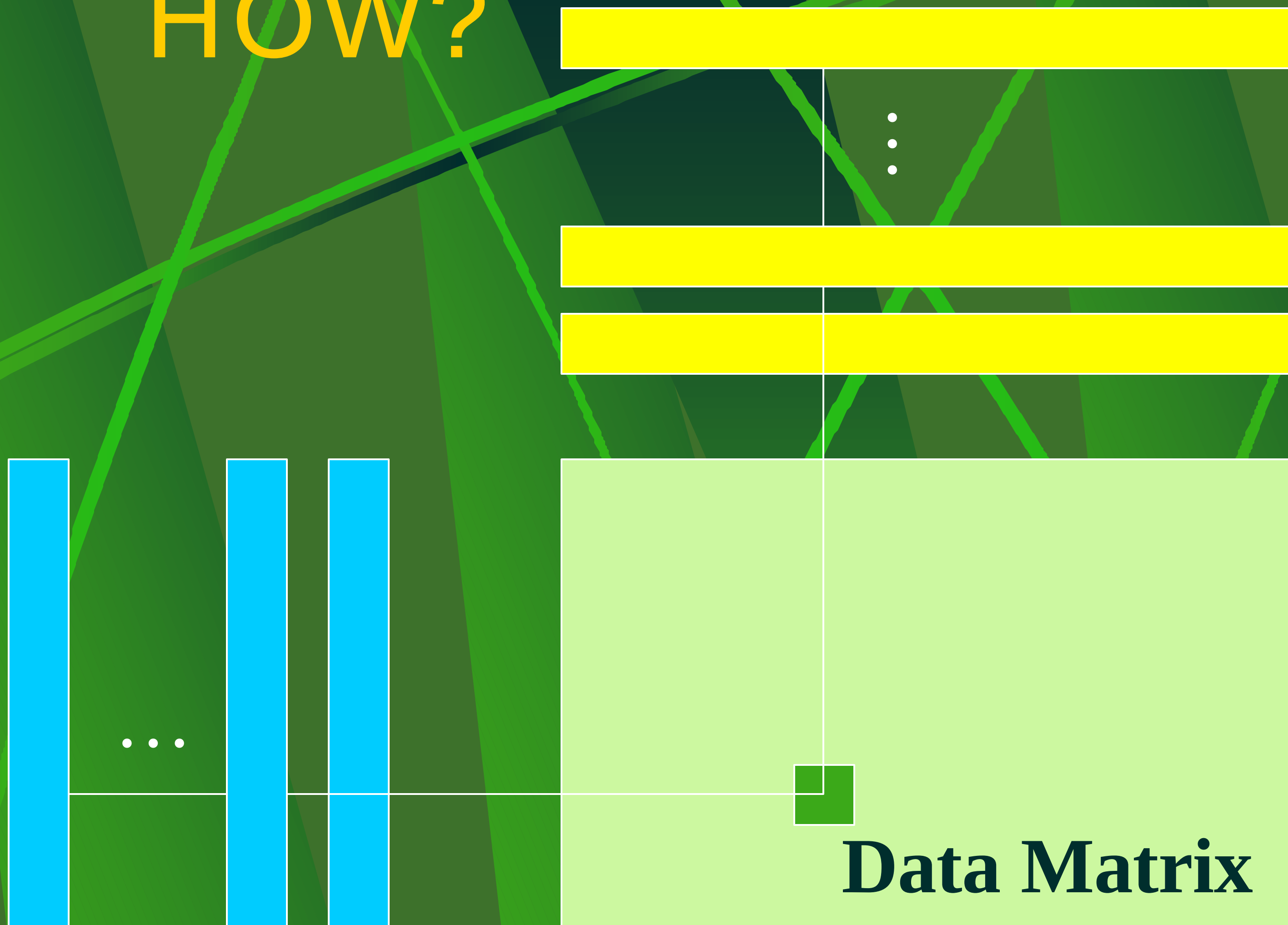
HOW?

Average concentration at each site

Variation across the region from day to day



HOW?



Linear Additivity

- The idea is to find a series of vectors such that their products sum to reproduce the data. This is the concept of linear additivity

Linear Additivity

Thus, we assume that the measured $PM_{2.5}$ mass can be written as contributions from p independent processes that affect that property across the spatial domain of the data.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}$$

Where i = 1,…, n days, j = 1,…, m locations and k = 1,…, p causal factors. f_{kj} involves the characteristics of the causal factor while g_{jk} provides the strength of the factor for the particular sample. e_{ij} is residual.

Linear Additivity

The nature of the causal factors will be discussed in more detail later. We can think of them in terms of emissions, dispersion, and transport over the space defined by the monitoring network

Factor Analysis

The equation can be rewritten in matrix form as

$$X = GF$$

The question is then how these two matrices, F and G, can be derived from the data.

Factor Analysis

Most factor analysis has been based on an eigenvector analysis. In an eigenvector analysis, it can be shown [Lawson and Hanson, 1974; Malinowski, 1991] that the equation estimates **X** in the least-squares sense that it gives the lowest possible value for

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(e_{ij}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}\right)^2$$

Factor Analysis

We explicitly acknowledge the problem to a least-squares problem and reformulate the problem so that we solve it by the minimization of the residuals weighted by estimates of the measurement uncertainty.

Factor Analysis

The Objective Function, Q, is defined by

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{s_{ij}} \right]^2$$

where $ó_{ij}$ is an estimate of the uncertainty in x_{ij}

Factor Analysis

This approach provides a more flexible framework for the analysis.

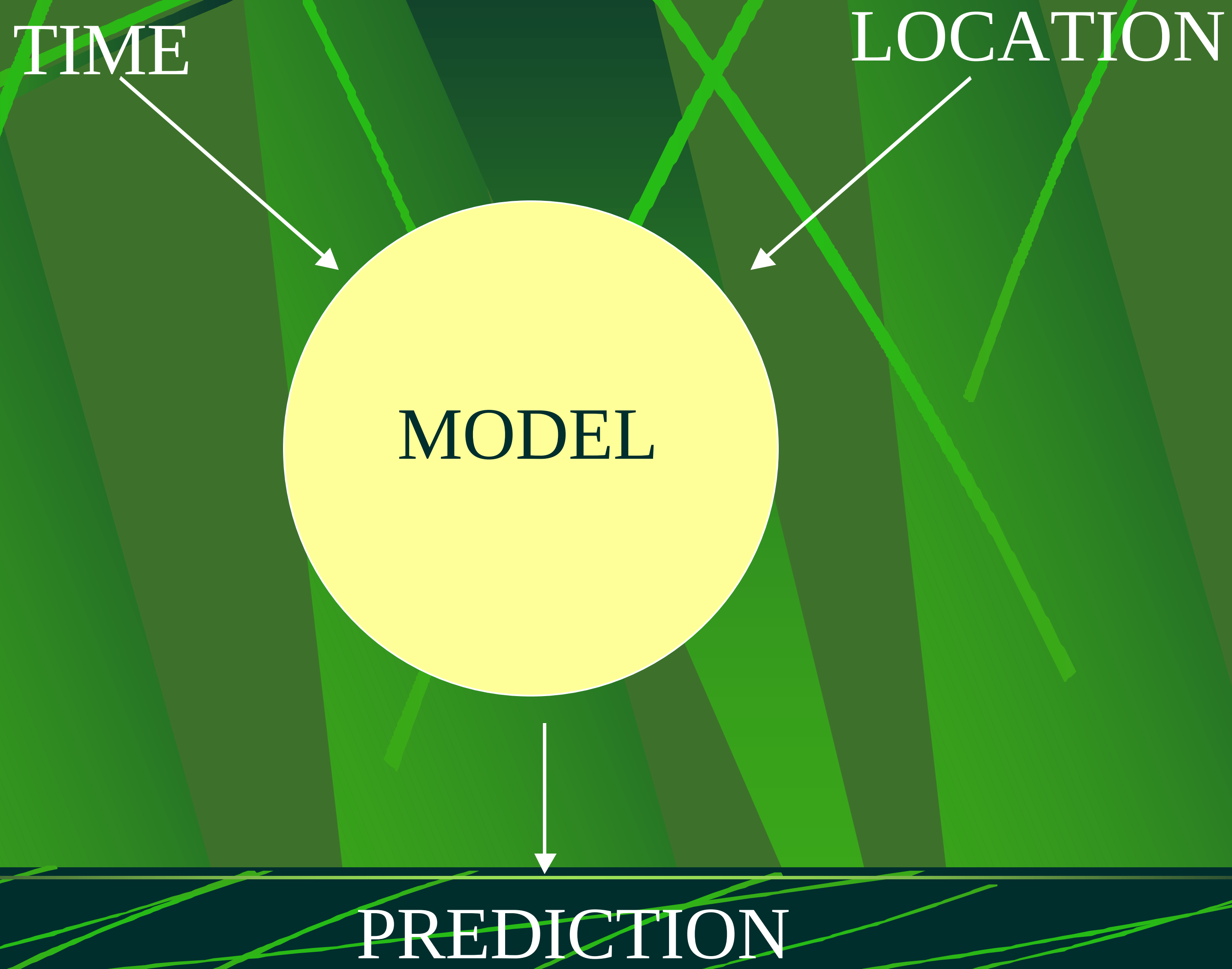
We can better separate signal from noise

We can develop more complex models that incorporate additional information that we have about the system

Factor Analysis

This additional information leads to "Advanced Factor Analysis" models

Statistical Modeling



Physical-Chemical Modeling

Temperature

Humidity

Gas Concentrations

Solar Radiation

Atmospheric Pressure

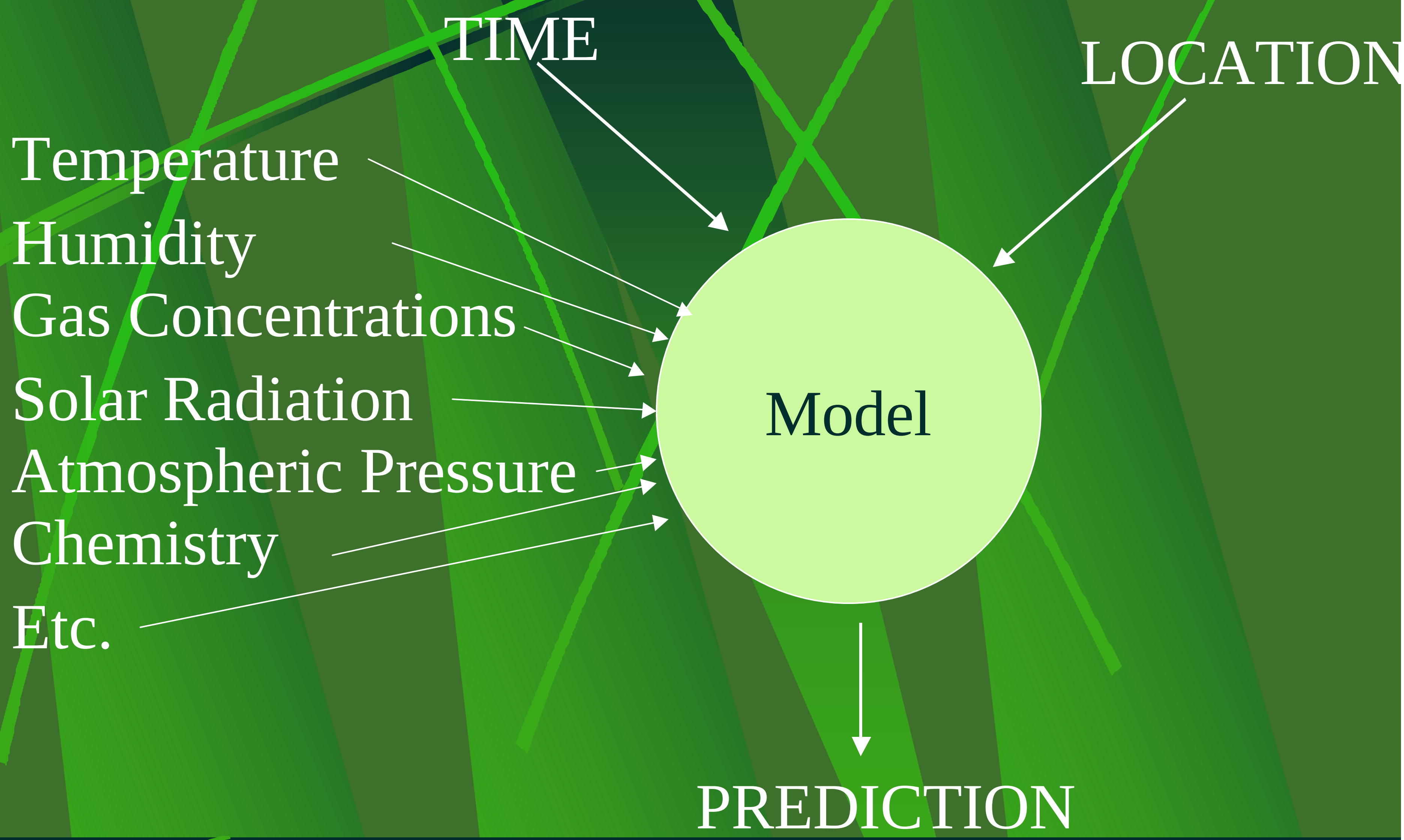
Chemistry

Etc.

MODEL

Prediction

Advanced Factor Analysis



Advanced Factor Analysis

For the advanced modeling, it is also assumed that the measured property is related to other data that are available

- Wind Speed and Direction
- Other pollutants like ozone
- Relative Humidity
- Temperature
- Etc.

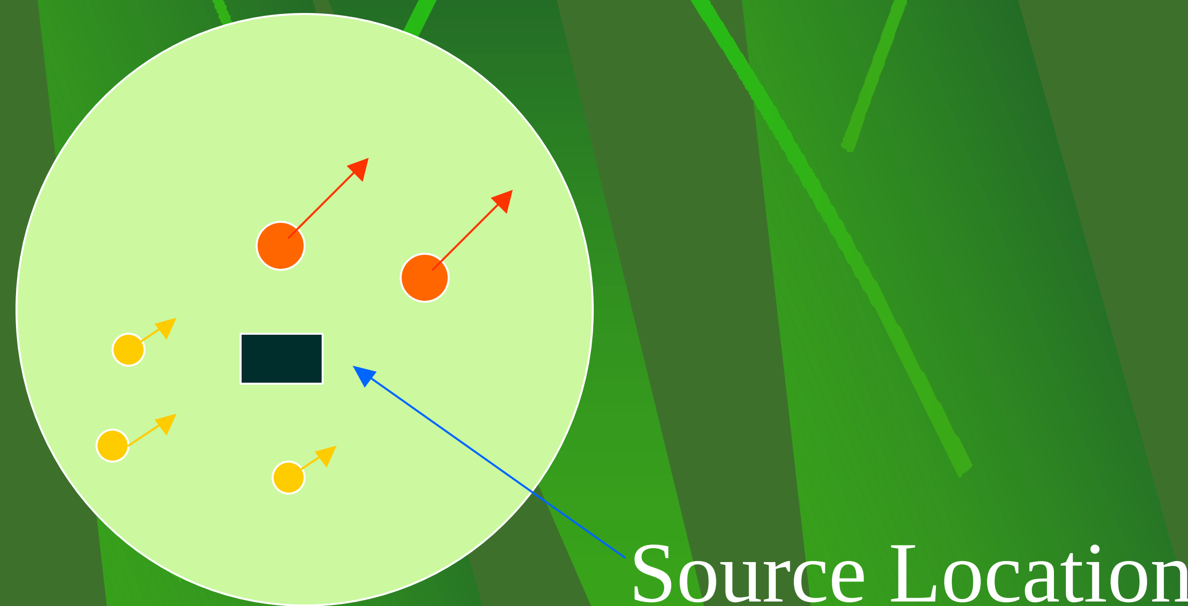
Advanced Factor Analysis

Combining wind information with the mass concentrations permits us to think in terms of $PM_{2.5}$ **flux.**

- Flux is the amount of $PM_{2.5}$ moving in a given direction.
- Flux is a vector so it has magnitude and direction.
- Average flux would be a function of concentration and wind speed and direction

Advanced Factor Analysis

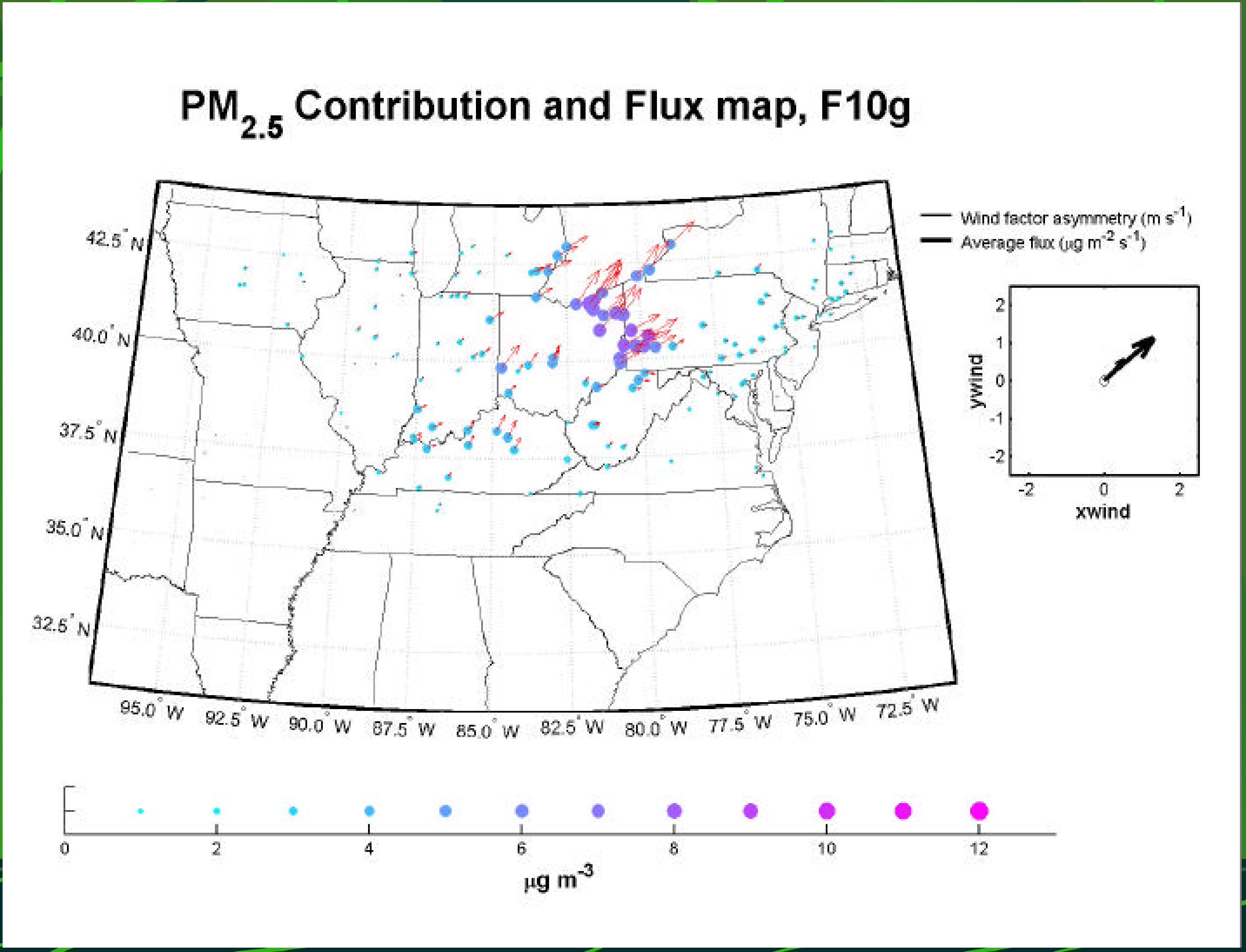
Suppose we can calculate the flux at various points within a given area.



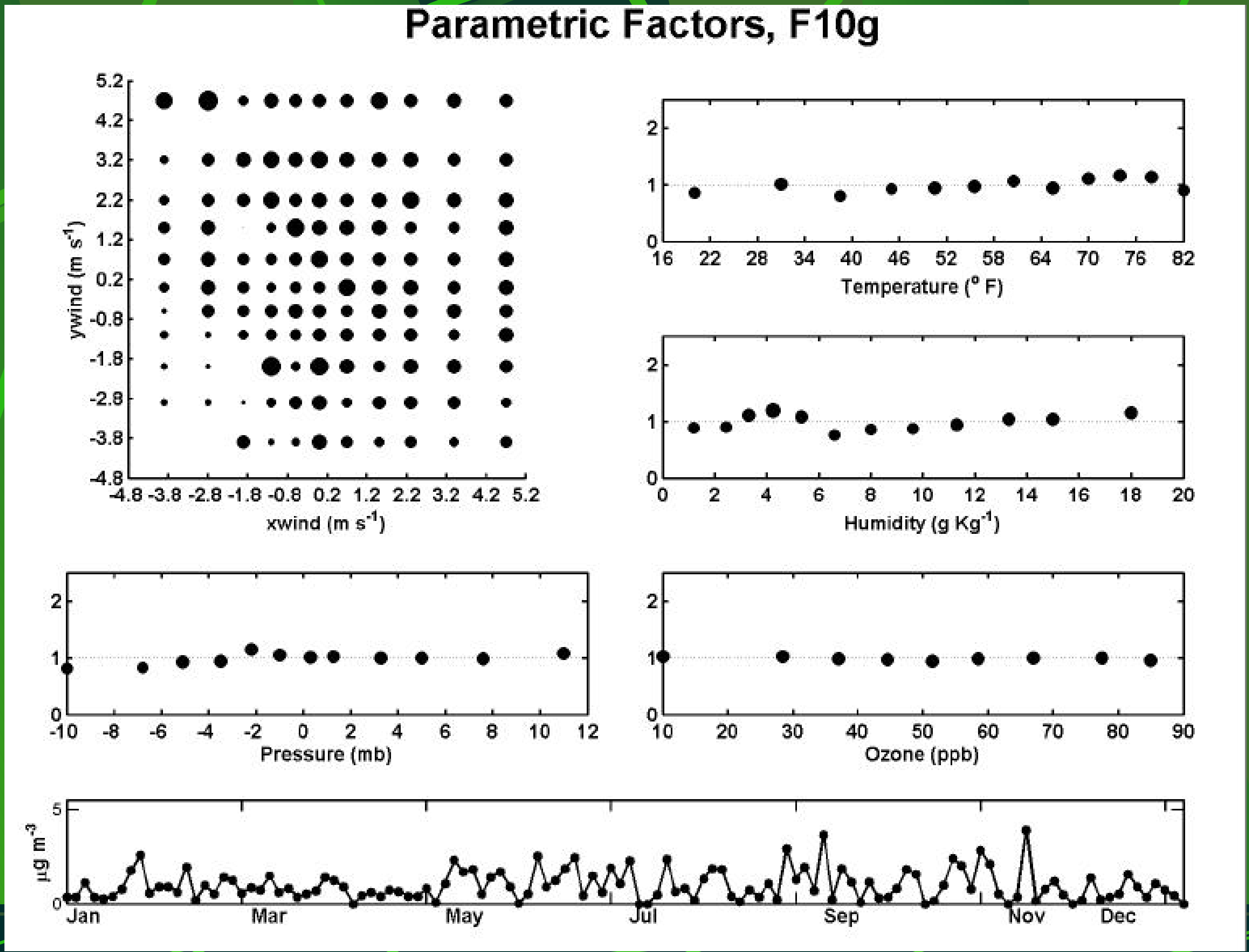
Advanced Factor Analysis

- The details of how such calculations can be done will be presented by Dr. Paatero in tomorrow's lecture.
- To illustrate what can be done, we present one of the factors that have been extracted in an advanced model from the 2000 $PM_{2.5}$ mass concentration data.

Advanced Factor Analysis



Advanced Factor Analysis



Advanced Factor Analysis

This is an on-going exploration of the models since we can incorporate additional information as appropriate.

QUESTIONS?